SAN DIEGO STATE UNIVERSITY

Antoni Luque Santolaria <aluque@sdsu.edu>

Small viruses

Simon Roux <sroux@lbl.gov> Thu, Dec 10, 2020 at 9:06 PM
To: Antoni Luque Santolaria <aluque@sdsu.edu>

This sounds good, I think you just got the set of sequences from Stephen, I will look from my side but the sequences I was thinking of should all be in Stephen's set already. The only difference is that Stephen's set is all the sequences predicted as complete, while we could also look at the ones predicted as "near-complete", but given that his set is already 7,000+ sequences, it seems like a good place to start.

One thing I should be able to add is to identify MCPs on these sequences, but I may not be able to get to it until ~ the end of next week, so maybe you'll be able to do this basic annotation much faster. Just let me know what makes most sense for you, and then we'll be happy to discuss the results you'll get !

Best,
Simon

On Dec 10 2020, at 12:13 pm, Antoni Luque Santolaria <aluque@sdsu.edu> wrote:

> Hi Simon,
>
> I think that what you propose would be ideal. If I have access to the (near-)complete phage genomes (with at least one major capsid protein), I will use the data as follows:
> 1. Study the distribution of Duplodnaviria capsids across ecosystems. Our working hypothesis is that there should be a shift towards smaller capsids as temperature increases. I believe that this analysis could lead to an excellent publication.
> 2. Incorporate the database to our training set to relate major capsid protein sequence and capsid architecture using random forest (the capsid architecture is obtained initially from the genome length). We have a paper in prep on this.
> 3. Investigate the molecular differences between MCPs across ecosystems. This could be studied within the first project listed above or as a separate project, depending on the findings and analysis level. Ideally, I would like to look at the folded structure in addition to the amino acid composition.
> Of course, you, Stephan, and any other folks from the IMG group involved in this dataset could be co-authors in the publications derived from the projects listed above.
>
> Regarding the annotation, we currently have two tools that could be used to annotate the associated capsid of a viral contig. They work for Duplodnaviria, but we plan to expand the same approach to other realms. That's part of my ongoing NSF Math Bio Award:
> 1. A linear regression model in log space predicts the capsid architecture from genome length.
> 2. A random forest predicts the capsid architecture from the major capsid protein sequence's features (amino acid frequency for the most part at this point).
> One question is if there are any specific set of standards that we should keep in mind to code these statistical approaches, so they can be integrated and maintained in your platform (for example, organization of python scripts and complementary files with specific input/output flows). Maybe, this process could occur naturally as the projects listed above move forward. Having some additional publications would definitely help to strengthen these approaches in the eyes of the community.
>
> Anyway, it's exciting that you got interested in our work. Let me know what would be the best way to access the data.

Best,
Toni

On Thu, Dec 10, 2020 at 12:22 PM Simon Roux <sroux@lbl.gov> wrote:

> Hi Toni,
>
> I fully agree, the dataset of (predicted) complete phage genomes from metagenomes that Stephen has gathered seems especially relevant and timely to your study, fingers crossed that at least some of these short circular contigs are indeed novel small-capsid phages !
>
> For the structural gene modules, we are always looking for ways to enrich the annotation as part of IMG/VR, and having a prediction of capsid properties / 3D rendering would be great. One of the main appeals on our database is the scale and the diversity of environments we incorporate, which typically gives you access to a very broad diversity of phages (i.e. we are less biased towards human gut than most other databases). I would be very curious to look at ecological distribution patterns for different capsid architectures (and also: which environment have the most phages with unusual/unpredictable capsids), as well as residue conservation across major capsid proteins.
>
> As to how to best move forward, I guess it depends somewhat on the current status of your capsid annotation pipeline. Maybe one first step could be for us to provide the sequence of (near-)complete phages we have in which at least 1 major capsid protein gene is predicted ? That could be a good dataset for you to check genome length vs capsid type relationship, while at the same time we would also get an estimate of how many (near-)complete phages have no major capsid protein annotated currently ?
>
> Best,
> Simon
>
> On Dec 10 2020, at 7:37 am, Antoni Luque Santolaria <aluque@sdsu.edu> wrote:
>
> > Hi Simon,
> >
> > Thank you for bringing in Stephen Nayfach and Nikos Kyrpides in the other email about short contigs. I believe that there is a lot to learn from those contigs. The CheckV tool seems very timely for us.
> >
> > Regarding the structural gene modules, my approach is to incorporate the biophysical constraints of capsids to assess the certainty of predicted structural genes further. One of my goals regarding annotation is to provide a way for the community to predict the capsid associated with uncultured viruses. It would be the equivalent of the genome architecture that is now common in the analysis of contigs. It would include a 3D rendering, capsid dimensions, and molecular weight, among other biophysical properties. Learning how to develop such a tool to be incorporated in platforms like IMG would great. If you think that this would be valuable and have any advice on how to move in this direction, your insight would be greatly appreciated.
> >
> > Best,
> > Toni

On Wed, Dec 9, 2020 at 12:05 PM Simon Roux <sroux@lbl.gov> wrote:

> Hi Toni,
>
> This sounds really interesting, and I would argue IMG/VR data are a great fit for this kind of analysis.
> For the IMG/VR website: if you over on "Search UViGs", you should see a "By UViG Attributes" menu item.
> This is where you will get a form that lets you apply multiple filters, including the ones I was mentioning: size (note it has to be > 0 , i.e. your minimum size can't be < 1), (estimated) completeness, and taxonomy.
>
> I would add that improving annotation of structural gene modules is definitely on our radar, so if it makes sense, we would be very happy to collaborate/contribute, especially to your analysis of IMG data. Let me know if you think it would make sense at any point.
>
> Best,
> Simon
>
> On Dec 9 2020, at 6:49 am, Antoni Luque Santolaria <aluque@sdsu.edu> wrote:
>
> > Hi Simon,
> >
> > Thanks for your quick response. I shared my paper with Forest and Bas, and it looks like you got contacted from both ends.
> >
> > I followed the link that you shared, but I'm not sure how to add the filtering options (see a screenshot of what I see). I navigated the options of the portal. I can narrow each filter separately through different tabs, but I couldn't figure out how to combine them. I'm sure I'm missing something trivial. Any advice on this?
> >
> > In any case, I have the paper that you recently published in NAR in my must-reads to learn how to navigate the system. Working with the IMG/VR is top of my list, even if the fragments are not complete. My plan next year, in fact, is to use the IMG data in combination with another approach that we are developing to predict the capsid architecture of tailed phages directly from the major capsid protein sequence. This would circumvent the requirement of having complete genomes.
> >
> > My lab aims to eventually provide computational models to generate high-resolution molecular capsid predictions from sequenced data. These so far are our baby steps.
> >
> > Thanks,
> > Toni
> >
> > On Tue, Dec 8, 2020 at 4:14 PM Simon Roux <sroux@lbl.gov> wrote:
> >
> > > Hi Forest & all,
> > >
> > > We do have a number of sequences that may possibly be relevant in IMG/VR ?
> > > Searching the whole database with the following criteria (https://img.jgi.doe.gov/cgi-bin/vr/main.cgi?section=ViralSearch&option=uvig):
> > >
> > > - size between 1 and 10,000 bp

- completeness between 80 and 100%
- taxonomic classification: Duplodnaviria

returns 97 sequences. Of course it's 97 out of 2.3 Millions, so a number of them (all ?) may just be crappy assembly and other artifacts. But maybe there are a few that would match your predicted new viruses in there ? Also note that the taxonomic classification is based on matches to the VOGdb, i.e. there are more sequences that fit the size/completeness criteria but are not classified as anything.

Toni: if you want to explore this further and have any question about IMG/VR, feel free to email me :-)

Best,
Simon

On Dec 8 2020, at 12:54 pm, Forest Rohwer <frohwer@gmail.com> wrote:

> Dear Virus Hunders,
> Toni Luque has predicted a set of small tailed phage based on the geometry of the capsids. He thinks they might be in hotter environments, but they could be anywhere. Have any of you seen any evidence of them (maybe in Minion datasets)?
> Sincerely,
> FLR

--
Dr. Antoni Luque
http://luquelab.com
Google Scholar profile
Assistant Professor
Department of Mathematics and Statistics
Viral Information Institute
Computational Science Research Center
San Diego State University

--
Dr. Antoni Luque
http://luquelab.com
Google Scholar profile
Assistant Professor
Department of Mathematics and Statistics
Viral Information Institute
Computational Science Research Center
San Diego State University

--
Dr. Antoni Luque
http://luquelab.com
Google Scholar profile
Assistant Professor
Department of Mathematics and Statistics
Viral Information Institute
Computational Science Research Center
San Diego State University